DOCUMENT RESUME

ED 322 150 TM 014 657

AUTHOR Mehrens, William A.; Green, Donald Ross
TITLE Standardized Tests and School Curricula.
PUB DATE Dec 86
NOTE 21p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; *Curriculum Evaluation; *Educational Assessment; Elementary Secondary Education; Local Norms; National Norms; *National Programs; School Districts; Scores; *Standardized Tests; Testing Problems; *Testing Programs; Test Interpretation; Test Use
IDENTIFIERS *Curriculum Based Assessment

ABSTRACT

This paper discusses the relationship of the content of nationally standardized and normed achievement tests and that of local school curricula and the effect that relationship has on the meanings and uses of the test scores. The following questions are considered: (1) whether tests have to match what is taught to be useful; (2) whether it is fair to teachers and students to use tests that include material not taught; (3) whether it is fair to use tests that do not include material that is taught; (4) whether local tests should be used instead of national tests; (5) whether national norms are needed; (6) whether some testing problems can be solved by the use of an item bank; and (7) whether or not it is possible to create a test that is tailor-made (customized) for local curricula and still obtain national norms by embedding nationally calibrated items. Five general issues are considered: inferences that can be made from standardized test scores; the basis upon which standardized achievement tests are considered valid; whether standardized achievement tests measure curricula; what the new technologies can do; and what schools can do. The respective answers to the seven questions are as follows: (1) no; (2) yes, provided the material is part of the target domain; (3) yes; (4) only for purposes where norms are not needed or where local norms (internal comparisons) are sufficient; (5) for most uses of test scores, national norms are the most easily understood and most informative type of norms--however, for some uses such as checking on just what the teachers are teaching, criterion reference scores may suffice; and (6) and (7) yes, in that more than one test can be blended together and put in one booklet--and no, in that unless the local curriculum very closely represents the domain sampled by a nationally normed test, two different tests are needed. Unless a very long test is acceptable, losses in reliability, domain coverage, and validity will ensue. (RLC)

STANDARDIZED TESTS AND SCHOOL CURRICULA

by

William A. Mehrens

Michigan State University

Donald Ross Green

CTB/McGraw-Hill

December, 1986

I. Introduction

The purpose of this paper is to discuss the relationship of the content of nationally standardized and normed achievement tests and that of local school curricula and the effect that relationship has on the meanings and uses of the test scores. Some of the questions considered are :

* Must tests match what was taught to be useful?
* Is it fair to teachers and students to use tests that include material not taught?
* Is it fair to use tests that do not include material that is taught?
* Shouldn't one use a local test instead?
* Are national norms needed?
* Can't these problems be solved by the use of an item bank?
* Isn't it possible to create a test tailor-made ("customized") for local curricula and still get national norms by embedding nationally calibrated items?

We will deal with these and other similar questions by discussing five general issues:

1. What inferences do we wish to make from standardized test scores?
2. On what basis are standardized achievement tests held to be valid?
3. Do standardized tests measure curricula?
4. The new technology: what can it do?
5. What are schools to do?

II. What Inferences Do We Wish To Make?

Validity refers to the accuracy of the inferences made from test scores. The only very direct inference we can make from any achievement test score is the degree to which a student knows the domain of material the test samples.

We may wish to make inferences about either current status and/or progress with respect to general and/or specific knowledge, skills, or abilities (KSA's). We may wish to make inferences about these KSA's in order to inform the individuals (or others) about their general level of functioning (or growth) in the field. Knowledge of this general level would be useful if an individual wished to have some notion about the likelihood of success in the next course (grade) or, if a terminal course, the adequacy of preparation for a job.

In order to make the general inference, the test needs to sample the general domain. For all of these cases, the degree of KSAs acquired typically can best be judged in a normative fashion. Given the mobility in our nation, national norms are most informative for these general inferences. We would add that the impression held by some that criterion-referenced scores are most useful for these sorts of purposes is faulty; while useful especially when the domains are small and finite, such scores have severe limitations and difficulties in interpretation without norms (Green, 1986; Mehrens & Lehmann, 1984).

At times, one desires to make inferences about more specific skills. If one wishes to diagnose and remediate specific weaknesses it is important to determine just what those specific weaknesses are. In training programs requiring specific skills, one desires to know whether the specific skills have been acquired. In these cases we

still almost always wish to generalize, but to a smaller domain.

At times we aggregate data about individuals because we wish to make inferences about the level of KSAs for a group. The group may, for example, be a class, a school building, or a school system. Here also we may be interested in current status and/or progress. Further, we could be interested in making general or specific inferences. In our view the far more common inferences either for individuals or groups are the general ones (see Green, 1983; Mehrens, 1984).

For purposes of curriculum evaluation for example, Cronbach stated better than 20 years ago that: "An ideal evaluation might include all the types of proficiency that might reasonably be desired in the area in question" (1963, p. 680). That is, the inference one is interested in making is the degree to which the group has mastered the total curricular domain. If the instructional domain were smaller than the curricular domain, and if the evaluation were focused only on what was specifically taught, there would be no data to tell us how the group functioned on the total domain of interest.

However we have considerable sympathy for the notion that instructional adequacy should be judged by measuring the outcomes related to the instructional objectives. We should not be judged instructionally incompetent if we did not teach, and our students do not know, the Roman Numerals. Nevertheless, even if we are interested in making inferences solely about the adequacy of instruction, we must be interested in whether or not the instruction resulted in the students being able to transfer the KSAs to new situations. Surely all teachers have transfer as an objective even if it is not one of those specific behavioral objectives written down in their lesson plan books. The issue is how general an inference do we wish to draw from a test score

covering the instructional domain?

Some individuals argue that if one infers to a general domain the inference is about amount of aptitude, not amount of achievement; and that educators should not be interested in making inferences about aptitude. At some level of generality the first part of the sentence is at least partly true but we disagree with the second part. Scholastic aptitude is commonly defined as developed abilities of the type that are related to subsequent school success. Educators should be in the business of helping students develop the general abilities useful for subsequent school success.

How closely one believes a test must be tied to a course of study to be an "achievement" test or how widely it must diverge to be called an "aptitude" test is a matter of opinion and convention. Current convention calls the tests in such batteries as the California Achievement Tests or the Stanford Achievement Test standardized achievement tests. The discussion in the next section related to their curricular validity justifies that label. One could also call the ACT and the Scholastic Aptitude Test achievement tests (see Jencks & Crouse, 1982) but we prefer to call them aptitude tests based upon their purpose (see Mehrens, 1982). Maintaining the distinction between aptitude and achievement tests according to their use is both appropriate and useful (Green, 1974).

III. Bases for Validity

For obvious reasons all publishers of nationally normed tests claim their tests are highly content valid. What they mean by this is that the content of the test is sufficiently representative of the domains that concern schools (e.g., reading comprehension, mathematics

computation, science, and so forth) so that one can legitimately draw inferences about students' KSAs in these domains from performance on the tests. Yet the several nationally normed test batteries differ in various ways including their content, sometimes to a surprising degree (Hoepfner, 1978).

Does this mean that some of these tests are not valid or is it possible that all of them may be simultaneously valid in spite of these differences? The latter seems to be the case; that is, each of them permits reasonable inferences about student KSAs with respect to the domains included in the battery; although the differences among these test batteries may be large, the procedures used in their development have enough in common to make it reasonable to say that the kinds of inferences possible are similar if not essentially the same.

The basic procedures followed by all publishers of major achievement batteries include the following steps in some form:

1. Select the domains to be measured. Essentially this means deciding what portions of the curricula found in schools are going to be represented in the battery. The major batteries currently available all have tests of reading, mathematics, and language arts, but only some have measures of science, social studies, and listening skills to mention just a few of the other possibilities.
2. Next define these content domains. For example, what is proper to include in a reading test? Should it include word analysis and/or word knowledge (vocabulary) subscores or be limited to comprehension?
3. Choose the sources to be used in setting up a content sampling procedure. All publishers look at curriculum guides from states

and school systems as well as the content of the textbooks in current use in the country. Several elements enter into the choices made at this juncture. One is the judgment about the importance of the materials at hand, e.g., is this textbook series more important than that one because it is more widely used? Another is the judgment about the conceptual merit of the materials. Foresight is also required; the test will be used from three to ten years after these plans are made and will not be useful if it samples a domain that no longer resembles school curricula.

4. The fourth step is to develop detailed specifications for each test based on careful analyses of the sources chosen. Thus the topics and objectives found at each grade level in the textbook series chosen as references will be detailed and compared with each other and with the sets of curriculum guides. The task is to select elements (topics, objectives) for representation and to write or select test items that measure them.

   The number of these elements is typically very large; for example, a few years ago a group of editors identified about 1500 different instructional objectives in reading basals and curriculum guides. Plainly only a small portion of them could be represented in a test of any reasonable length. Even though those chosen are usually restricted to the elements common to most of the materials examined, the number of possibilities remains large. Thus it happens that different tests will not have exactly the same content nor even sample the same domain. They are however likely to be similar because large portions of the content found in school curricula across the country are

very similar.

Properly done these steps lead to a test from which valid inferences about student KSA's can be drawn. Notice that no test can include all objectives found in the typical curriculum of a school system and so there will always be material taught but not tested. Furthermore it is likely that there will be material tested that is not taught in some classrooms. Neither of these circumstances in and of themselves invalidates the use of the test in those classrooms.

IV. Do Standardized Achievement Tests Measure Curricula?

Sample vs. Domain

Given the description in the previous section of how tests are constructed it is obvious that at some level of generality tests measure the curricula in U.S. schools. However, not all curricula are identical and not all tests sample exactly the same domain. In considering the match between any two tests, any two curricula, or between any test and a curriculum one must first consider the difference between matching samples and matching domains. One cannot infer a lack of match of the domains from a lack of match of the samples. Yet the common method of studying test/text overlap is to look at test items. Suppose, for example, we wished to see if two forms of a 60 item final exam were comparable and matched the content of Measurement and Evaluation in Education and Psychology (Mehrens and Lehmann, 1984). Suppose we classified the content domain of basic educational and psychological measurement into a 1260 cell matrix. Obviously if the matrix cells were mutually exclusive and each item only measured the content of a single cell, at most a 60 item test could cover only 60 of the cells in the matrix. The other form of the test may well

cover 60 different cells but both tests could be representative samples of the domain to which we wish to infer. The equivalence of the two tests would be determined, in part, by seeing how well they correlate. If the correlation were 0.90, then the reliability index (or correlation between either test score and the true, or domain, score) would be estimated to be 0.95. One could thus infer from either test to the domain with a fair degree of precision even though the two tests measured totally different cells in the 1260 matrix.

Kinds of Overlap

Writers discuss two kinds of overlap: content tested but not taught and content taught but not tested. As Mehrens pointed out:

> "Because no one really believes standardized tests measure all important educational outcomes or even an individual teacher's set of instructional objectives and no one makes an inference from a standardized test to either of those domains we need not worry much at all about the content not tested mismatch insofar as the inference from a score to a domain" (1984, p.11).

Porter is not so sanguine:

> "Substantive differences in content covered in textbooks and test along with differences among teachers in beliefs about what should be taught call the concept of a national curriculum into question.... These differences promote diversity in what is taught, rather than consensus, and give rise to conditions in which standardized tests may consistently underestimate student achievement" (Porter, as quoted in Captrends, 1985)

We have trouble believing that anyone would infer that the only things a student knows about a subject matter are the items he/she

answered correctly in a test or the percent of such items known in the domain sampled. Obviously a student has knowledge in a subject matter outside the domain sampled by a standardized achievement test. The test score is neither an overestimate nor an underestimate of that knowledge. It is not an estimate of that knowledge at all.

Content tested but not taught also may exist. Whether that unfairly lowers an individual's (or group's) norm referenced score depends on how the amount of that content compares to the norm group's distribution of untaught content -- something we never know. This amount is likely to be large only in those cases where the local curriculum is sharply deviant from those of other school systems.

How Overlap Gets Measured

Curriculum overlap is typically measured by asking judges to match the test items with the written curriculum. Instructional overlap is measured by matching the test items with the actual instruction. The approach we prefer for determining whether there is content tested that is not taught is to determine, for each item, what objective the item tests and then go to the curricular materials (cumulated over previous grades) to see if the objective had been covered. In essence the question is as follows: If the students understood the concepts covered in the curricular materials should they be able to answer the item in the test?

Degree of Overlap

The degree of overlap between test and curriculum varies depending on subject matter, methodology, and the particular test and curriculum being compared. Estimates range as low as 54.2% (Freeman et al, 1980)

up to 100% (Bower, 1982). The research producing the lower figure cited here used a very specific classification scheme and matched against a one grade text rather than a cumulative curriculum.

Impact of Different Levels of Overlap

Phillips and Mehrens (1985, 1986, in press) and Mehrens and Phillips (1985, 1986a, 1986b) have conducted a series of studies which show that different textbook series and informal curricula within a single district have no significant impact on test total, objective or item scores.

This is as it should be. For the inferences to be correct, the test results should not be strongly influenced by specific and generally small differences in content coverage. This does not mean that standardized test scores are insensitive to instruction. The research was not designed to study the impact of instruction vs. no instruction; nor was it designed to study the impact of two totally different curricula--one which covered the objectives tested and one which did not. It does seem likely that differences would be found in those types of studies since the tests should reflect the degree to which students, in general, have acquired knowledge of the domain sampled. However the point of the Mehrens and Phillips studies is that the naturally occurring differences in curricula within a school district (due to either teacher emphasis or textbook differences) are not great enough to be observable on the results of the standardized tests.

V. The New Technology: What Can It Do?

Ways to address the concerns regarding content match that we have been describing have been proposed in recent years. Many of them are

based on the use of item banks and item response theory (IRT) which together make it possible, under certain circumstances, to create "customized" tests which will not only measure the local curriculum but also yield estimates of performance on a scale defined by a nationally normed test. Within certain limits these procedures do work, and they do make it appear that by using item banks of calibrated items most of the problems can be solved. However there are many pitfalls (Green, 1985).

Adequate estimates of performance on a normed scale can only be obtained from a set of items that, like the originally normed test, is a reasonable representation of the trait. In other words the inference has to be to the same domain as that permitted by the original test. It follows that the customized test may either be really two tests or that the effort was unnecessary. In the first instance the items representing the normed trait constitute one test, and the items representing the local curriculum represent the other. Putting the two together may have administrative advantages but there are really two tests and two sorts of scores even if the two tests overlap in content to some extent and appear in a single test format.

Of course if the set of items that adequately represents the normed trait also adequately represents the local curriculum, there is only one test and one score. However in that instance there was no need to create the "customized" test.

It may be noted that in the "two test" instance there are a number of possible misunderstandings that may arise. For example, one may calibrate the items referenced to the local curriculum on the scale of the nationally normed test but after a period of time (i.e., after instruction) those calibrations may no longer be accurate (Yen, Green,

& Burket, 1986).

Still, it follows that all these things can be done and school systems can arrange for custom built tests which precisely fit their curriculum and which via the route of embedded calibrated items (which may, and usually should in some part, reference content not part of their curriculum) provide estimated national norms. Clearly there are some important trade offs to be examined when considering following these procedures in a district testing program. Probably the most important element in making decisions about how to proceed is having a clear understanding on the part of all concerned - teachers and administrators alike - about what are the necessary and appropriate purposes to be served by the program. Programs that are merel a compromise between conflicting parties with differing goals are not likely to satisfy anyone.

VI. What Are Schools To Do?

Schools should administer standardized achievement tests to their students. External norm referenced test data are essential for students, parents, and significant others in judging the status and progress of individual students. The aggregation of such data is of value in assessing the local school's curriculum. In choosing tests, educators should look at the content of the various achievement tests. It is probably somewhat self-serving but wise to choose a test whose content samples a domain similar to the curricular domain of the school. It is likely that no test will appear to have a perfect match for every subject for every grade level. It is also likely that no reasonably popular test will depart too drastically from the curricular domain of a school. If a school's curriculum is not very related to

any standardized test's content domain "that local curriculum may indeed be too bizarre and in need of some changes" (Mehrens, 1984, p.13).

Once a test is chosen it is appropriate to let all the educators see a copy of it. It is appropriate for the curriculum committee to look for test-curriculum mismatch to determine whether the curriculum should be broadened as long as the domain is made larger rather than just adding on some specific objectives or particular item types. It is not appropriate to narrow a curriculum based on test coverage. The taught but not tested mismatch will and should be present. No test covers all that could or should be taught.

The research on the impact of different levels of overlap previously cited suggests that even schools that stress and publicize test results may not need to fear too much that their schools' scores will be significantly influenced due to differential curriculum-test match. Likewise schools need not fear too much that teachers will inappropriately teach to the test. (We realize there will be exceptions. Unprofessional educators exist.) Teachers surely realize that the tests only sample the important objectives of the school and they realize they are to teach to the domain of the objectives, not the particular sample tested and certainly not to the specific questions.

Inferences that schools make from the test scores should be to the domain the test samples--not some other domain. Instructional effectiveness cannot be inferred if the test domain does not match the instructional domain. Even if the domains do match, any inference about instructional effectiveness is a weak inference and appropriate research designs and/or statistical adjustments would have to be implemented prior to any such inference (see Haertel, 1986).

As we have suggested earlier in this article, there are times when the inference of interest is to the more narrow domain of what is specifically covered in the local curriculum. This is the case when one wishes to make inferences regarding the adequacy of the instruction, not the adequacy of the curriculum. Keeping in mind that even the measurement of local instructional objectives would demand the testing of understanding and therefore the testing of material not directly taught, it may well be useful to build tests specifically over the domain of interest for such inferences. Prior to the recent popularity of criterion referenced testing most school districts relied on teacher made tests to do that job. In fact, many of those tests were criterion referenced with respect to the inference made from them. However, teachers' tests usually are not constructed in ways that permit inferences to be made to the full range of local objectives. Thus, many school districts currently construct tests to precisely measure the local objectives. Such tests are frequently built by trained personnel in the district's office of evaluation. They can also be built by an external contractor. Testing companies have been doing an increased amount of such test construction in recent years.

Such tests are likely to pro ide useful information over and above what one can infer from standardized tests. (Such testing should not be viewed as a replacement for the standardized tests.) However, there are reasonable questions about the conditions necessary in the local district for such tests to be maximally useful and whether or not they are ever likely to be cost effective. For district tests (as opposed to individual teacher built tests) to be useful, it would be necessary to have the same curriculum in place in all the classrooms

that use the test. If there is not a fairly comprehensive and rigid district curriculum, such tests would not be any more fair for purposes of instructional evaluation to the school buildings (or individual teachers) who depart from the domain the test samples than would a standardized achievement test. A requirement for such tests to be cost effective is that decisions get made on the basis of results. Recall that the only direct inference one can make from a test is the degree to which a student knows the material that the test samples. Any inference about why the students know the material to that degree is a weaker inference.

Thus, the major purpose of such locally developed tests should not be for purposes of teacher evaluation, but rather for the purpose of guiding teachers in the specific decisions they make regarding what and how to teach the students. Whether the expenditure of the resources to build such tests is more cost effective than expending the same amount of resources in training teachers to build more effective teacher made tests is something reasonable experts could debate. We are inclined to think it is for those subject matters where the district wide curriculum is well developed and communicated to the teachers with the expectation that it be followed.

VII. Conclusions

At the beginning of this paper we posed some questions. We hope those who have read this far know what our answers to them are, but to remove any doubts they will be given here:

*Must tests match what was taught to be useful?

For most of the purposes for which standardized tests are

used the answer is: No, the two should be logically and closely related and should overlap to some degree but they need not and usually should not match perfectly.

*Is it fair to teachers and students to use tests that include material not taught?

Yes, provided this material is a part of the target domain.

*Is it fair to use tests that do not include material taught?

Yes. No test of reasonable length could test everything taught even if it were all measurable in paper and pencil form.

rShouldn't one use a local test instead?

Only for those purposes where norms are not needed or where local norms (internal comparisons) are sufficient. National norms are generally needed. Criterion referenced scores taken alone have severe limitations.

*Are national norms needed?

For most uses of test scores national norms are the most easily understood and most informative. However for some uses such as checking on just what the teachers are teaching, criterion referenced scores may suffice.

*Doesn't current technology such as item banks and Item Response Theory make it possible to build tests that serve ALL purposes simultaneously?

Yes and no. Yes in that more than one test can be blended together and put in one booklet. No in that (a) unless the local curriculum represents the domain sampled by a nationally normed test very closely, two different tests are needed, and (b) unless a very long test is acceptable,

losses in reliability, in domain coverage and thus in validity will ensue.

BIBLIOGRAPHY

Bower, R. (1982) Matching Standardized Achievement Test Items to Local Curriculum Objectives. Symposium paper presented at the annual meeting of the National Council on Measurement in Education, New York City.

Cronbach, L. J. (1963). Evaluation for Course Improvement. Teachers College Record, 64, 672 - 683.

Freeman, D. J., Kuhs, T. M., Porter, A. C., Knappen, L. B., Floden, R. E., Schmidt, W. H. & Schwille, J. R. (1980). The fourth grade mathematics curriculum as inferred from textbooks and tests (Research Series No. 82). Fort Lansing: Michigan State University, Institute for Research on Teaching.

Green, Donald Ross (1974). The distinction: some conclusions. In D.R. Green (Ed.), The aptitude-achievement distinction (pp. 349-354). Monterey, CA: CTB/McGraw-Hill.

Green, Donald Ross (1983, April). Content validity of standardized achievement tests and test curriculum overlap. Paper presented at the annual meeting of the National Council and Measurement in Education, Montreal.

Green, D. R. (1985, April). Misinterpreting and misusing tests: some ways. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Green, D. R. (1986). Interpreting scores from standardized achievement tests. National Association of Secondary School Principals Bulletin. (In Press).

Haertel, Edward (1986). The valid use of student performance measures for teacher evaluation. Educational Evaluation and Policy Analysis. Vol. 8, No. 1.

Hoepfner, R. (1978). Achievement test selection for program evaluation in Wargo, N. J. & Green, D. R. Achievement testing of disadvantaged and minority students for educational program evaluation. Monterey, CA: CTB/McGraw-Hill.

Jencks, C. & Crouse, J. (1982). Should we relabel the SAT... or replace it? in W. B. Schrader (ed.), New Directions for Testing and Measurement: Measurement, Guidance, and Program Improvement. No. 1 (pp. 33 -49). San Francisco: Jossey-Bass.

Mehrens, W.A. (1982). Comments on "Should we relabel the SAT" ... or replace it? in William B. Schrader (Ed.), New Directions for Testing and Measurement: Measurement, Guidance, and Program Improvement. No. 13 (pp 54-57). San Francisco: Jossey-Bass.

Mehrens, W. A. (1984). National tests and local curriculum: Match or mismatch? Educational Measurement: Issues and Practice, 3, 3, 9 - 15.

Mehrens, W. A. & Lehmann, I. J. (1984). Educational Measurement and Evaluation in Psychology and Education (3rd ed.). New York: Holt, Rinehart, and Winston.

Mehrens, William A. & Phillips, S. E. (1985, April). Sensitivity of ite statistics to curricular validity. Paper presented at the Nationa Council on Measurement in Education Annual Meeting, Chicago.

Mehrens, William A. & Phillips, S. E. (1986 a). Detecting impacts of cur ricular differences in achievement test data. Journal of Educational Measurement. 23, 3, 185 - 196.

Mehrens, William A. & Phillips, S. E. (1986b). Sensitivity of special group item statistics to curricular validity. Paper presented at the American Educational Research Association annual meeting, San Francisco.

Phillips, S. E. & Mehrens, William A. (1985, April). The effects of curricular differences on achievement test data at the item and objective level. Paper presented at the American Educational Research Association Annual Meeting, Chicago.

Phillips, S. E. & Mehrens, W. A. (1986, April). The effects of curricula differences on the achievement test scores of special groups. Pape presented at the American Educational Research Association Annual Meeting, San Francisco.

Phillips, S. E. & Mehrens, W. A. (in press). Curricular differences and unidimensionality of achievement test data: an exploratory analysis. Journal of Educational Measurement.

Porter, A. C. (1985) Do Tests and Textbooks Match?, Captrends, vol. 11, No. 1, 1 - 3. Northwest Regional Educational Laboratory.

Yen W.M., Green, D.R., & Burket, G.R. (1986, April). Valid Normative information from customized achievement Tests. Paper Presented at the annual meeting of the National Council on Measurement in Education, San Francisco.